

## 本周周报（2012. 10. 22-2012. 10. 28）

马晓红

本周工作：

1. 阅读了两篇关于 entity retrieval 的论文。

Keyword++: A Framework to Improve Keyword Search Over Entity Databases: 讲了在数据库上进行关键字查询的方法。具体的实现方法是首先根据关键字构造出 SQL 语言，然后根据 SQL 语言在数据库上进行检索。由关键字构造 SQL 语言是通过查分查询对(Differential Query Pair)来实现的。比如对于 small IBM laptop 这个查询，要知道 IBM 对应的 SQL 语句，先构造 small IBM laptop 和 small laptop 这两个查询在搜索引擎上进行搜索，然后根据两个结果的对比来判断 IBM 对应的 entity 的哪个属性，进而构造 SQL 语句。

EntityRank: Searching Entities Directly and Holistically: 这篇文章主要介绍了一种检索结果为 entity 的方法，着重于对于检索到的 entity 的有效排序。这种方法充分考虑了实体的上下文环境、检索的全面性、不确定性和相关性等问题。

2. 看了一些关于信息检索中语言模型。重点了解了一下语言模型中的平滑方法。我了解的平滑方法有：Jelinek-Mercer 法（线性插值）、Dirichlet prior 法、拉普拉斯光滑法、Absolute discount 方法。并且看到了有研究说明了对于长短不同的查询，这些方法的效果是不同的。对于短查询，效果 Dirichlet prior > Absolute discounting > Jelinek-Mercer；对于长查询，Jelinek-Mercer > Dirichlet prior > Absolute discounting。UIUC 的工作中的语言模型采用的是 Jelinek-Mercer 的光滑方法，但是还不知道具体的查询的长短，如果不是长查询我觉得可以从平滑方法上尝试改进。这只是一思路。

下周工作：

1. 某些课程要结束了，有一些相关的课程作业要完成。
2. 继续看关于 IR 方面的论文吧~